Rooting out the Rumor Culprit from Suspects

Wenxiang Dong*,Wenyi Zhang* and Chee Wei Tan†
*University of Science and Technology of China, Hefei, 230027, China
†City University of Hong Kong, 83 Tat Chee Avenue, Hong Kong
javin@mail.ustc.edu.cn, wenyizha@ustc.edu.cn and cheewtan@cityu.edu.hk

arXiv:1301.6312v2 [cs.SI] 5 Feb 2013

***Abstract*—In this paper, we study the problem of rooting out a single rumor source: supposing that a rumor, which originates from one node among a set of suspect nodes in a network, is spreading across the network following the susceptible-infected (SI) model, how can we and how well can we identify this rumor source? With the *a priori* knowledge of the set of suspect nodes and a snapshot observation of infected nodes, we construct a maximum *a posteriori* (MAP) estimator to identify the rumor source. For regular tree-type networks, we establish exact and asymptotic results on the detection probability of the source estimator. In particular, if the network is not linear, we have the following asymptotic results of the correct detection probability: First, when every infected node belongs to the suspect set, it grows from 0.25 to 0.307 as the node degree increases from three to infinity, a result first established in [1], [2] via a different approach. Second, when the suspect nodes form a connected subgraph of the network, it significantly exceeds the *a priori* probability, and reliable detection is achieved as the node degree becomes large. Third, when there are only two suspect nodes, it is at least 0.75 and increases with the distance between the two nodes. Our analysis leverages ideas from the Pólya's urn model and sheds insight into the behavior of the rumor spreading process.**

***Index Terms*—maximum a posteriori estimation, Pólya's urn model, rumor spreading, social network, susceptible-infectious model.**

I. Introduction

Spreading of epidemics, information, ideas and innovations through social networks is ubiquitous in the modern world [4], [3], such as propagation of diseases and immunization, information diffusion through cyberspaces, upsurge of hot research topics, and adoption of medical and agricultural innovations. In general, any of these situations can be modeled as a rumor spreading process through a network [1], [5]. In order to control, accelerate or prevent these spreadings, a key challenge is to identify the rumor source only based on the knowledge of network structure and the observation of infected nodes.

In many practical scenarios, there does exist some *a priori* knowledge that certain nodes are more likely to be the rumor source and the others are less likely so, and there are many situations that only a handful of nodes in the network have the potential to initiate a rumor spreading process. When a disease among human beings is spread from cities to cities, only the travellers from earlier infected cities may cause the epidemic outbreak in a new city. When a rumor arises in the blogspace, its origin can be the popular and eloquent bloggers. When a network of water pipes is polluted by microorganisms or chemical substances, only the key points need to be examined to identify the pollution source. When the cyberspace is hit by junk mails or computer viruses, most of the victims should not be suspected as the culprit. Therefore, we do not necessarily treat every infected node as possible rumor source when identifying the source. Hereafter, we call the nodes that have the potential to initiate a rumor spreading process across a network as *suspect nodes*.

In this paper, we consider the issue of identifying a rumor source from suspect nodes, conditioned on a snapshot observation of infected nodes in a network. Our goal is to identify the rumor source, only based on the knowledge of network structure, the set of suspect nodes and the observation of infected nodes.

A. Related Works

Over the past decades, abundant works have dealt with the problems on epidemic outbreaks across networks. The main attention has been paid to understanding the impacts of network structure and infection/cure rates on the diffusion processes, such as in [6], [7], [8], [9], [10]. Besides, many researchers have developed network inference techniques to learn the underlying network parameters and predict the propagation characteristics, such as in [11], [12], [13], [14]. In addition, the issue of extracting the influential source nodes for epidemic spreadings has also been considered, such as in [4], [15], [16]. Recently, the rumor source estimation problem has been formulated and studied for the first time.

The pioneering work, which tackled the estimation problem of a single rumor source using the susceptible-infectious (SI) model in [1], [2], has led to an upsurge of the topic on rumor source estimation. For regular tree-type networks, they constructed a maximum likelihood (ML) estimator and derived its asymptotic performance. When the node degree $\delta = 2$, i.e., for a line graph, the asymptotic correct detection probability is zero; when $\delta = 3$, it is 0.25; when $\delta > 3$, it is a positive constant value $\phi_1(\delta)$, which approaches 0.307 as δ grows large. Besides, other types of networks and models are also considered, such as geometric trees [1], random graphs [17], estimation of multiple rumor sources [18] and the susceptible-infectious-recovered (SIR) model [19], and estimation of a single rumor source using sparsely distributed monitors' measurements [20]. However, those works assume that every node in the network has the potential to initiate the rumor spreading process.

B. Our Contributions

In this paper, we treat the estimation problem of a single rumor source, with the *a priori* knowledge that the rumor source is restricted in a specified *suspect set* S of suspect nodes in a network G. The rumor spreading process across the network G is modeled by the SI model, which is a variant of the common SIR model for rumor spreading [5]. Here, it is reasonable to use the SI model, since we do not have much side information about the states of recovered nodes [1]. Conditioned on an observation of n infected nodes in the network G, we construct a maximum *a posteriori* (MAP) estimator to identify the rumor source. In order to study the performance of the estimator, we further analyze $\mathbf{P_c}(n)$, the correct detection probability upon observing n infected nodes. For regular tree-type networks with node degree δ, we establish exact and asymptotic detection results on $\mathbf{P_c}(n)$ under three interesting scenarios, assuming a uniform *a priori* distribution of the rumor source over S. The analysis leverages ideas from the Pólya's urn model [21] and sheds insight into the behavior of the rumor spreading process.

First, in the extreme case where S contains all nodes in the network G, i.e., every infected node has the potential to be the rumor source, our estimator reduces to that in [1], [2]. However, different from [1], [2], we obtain the same asymptotic detection results on $\mathbf{P_c}(n)$ as well. We conduct the argument via a rather different approach from theirs, only relying upon the discrete and limiting distribution of the number of infected nodes. By contrary, the authors in [1], [2] conduct the argument explicitly relying upon the exponential distribution of the infection time. Besides, we also obtain exact detection results on $\mathbf{P_c}(n)$ when the node degree $\delta = 2$ and $\delta = 3$, and our approach can be also used in the following cases.

Second, in the case where S with cardinality k forms a connected subgraph of the network G, which may correspond to a small social community. When the node degree $\delta = 2$ and $\delta = 3$, we can obtain both exact and asymptotic detection results on $\mathbf{P_c}(n)$. When $\delta > 3$, the asymptotic value $\lim_{n\to\infty} \mathbf{P_c}(n)$ is a positive constant value $\phi_2(\delta, k)$ for all $k \ll n$. Besides, $\mathbf{P_c}(n)$ significantly exceeds the *a priori* probability $1/k$ if the network is not linear (with $\delta \geq 3$), and the MAP estimator achieves reliable detection, i.e., $\lim_{n\to\infty} \mathbf{P_c}(n) \to 1$, as δ grows sufficiently large.

Third, in the case where S contains only two suspect nodes separated by their shortest path distance d, which may be a quantitative measure of the social closeness between them. Without loss of generality, consider $d < n$. For a line graph, we obtain exact detection results for $\mathbf{P_c}(n)$; when $\delta = 3$, $\lim_{n\to\infty} \mathbf{P_c}(n) = 0.75$ if $d = 1$ and $\lim_{n\to\infty} \mathbf{P_c}(n) \approx 0.886$ if $d = 2$; when $\delta > 3$, $\lim_{n\to\infty} \mathbf{P_c}(n)$ is also a positive constant value $\phi_3(\delta) > 0.75$ if $d = 1$, which further approaches one as δ grows sufficiently large. In addition, $\mathbf{P_c}(n)$ increases with d for general regular trees.

C. Organizations

First, Section II describes the framework of the SI rumor spreading model and the MAP estimator to identify the rumor source. Second, we establish theoretical results on the detection probability of the estimator for regular tree-type networks under three scenarios with the Pólyas urn model in Section III. Third, we carry out simulation experiments to study the detection performance of the estimator for regular trees in Section IV. At last, we conclude the paper in Section V.

II. Rumor Spreading Model and Rumor Source Estimator

In this section, we describe the SI rumor spreading model, define the MAP rumor source estimator for regular trees, and introduce the concept of local rumor center.

A. Rumor Spreading Model

In general, a undirected network $G = (V, E)$ consists of a set of nodes V and a set of edges E. V is assumed countably infinite so as to avoid any boundary effect, and any pair of nodes may infect each other if and only if they are connected by an edge in E. The network is assumed static, without nodes joining or leaving.

In this paper, the rumor spreading process is modeled by the susceptible-infectious (SI) model, a variant of the common susceptible-infectious-recovered (SIR) model for rumor spreading. In the SI model, it is assumed that the system operates in a progressive fashion, i.e., once a node gets infected it keeps the rumor forever.

We consider the case that only a single node in an *a priori* specified *suspect set* S can be the rumor source in each resulting infection across the network G, where $S = \{s_1, s_2, \cdots, s_k\} \subseteq V$ has cardinality k. We assume *a priori* distribution $\mathbf{P_s}$ of the rumor source over the nodes in S, and thus $\mathbf{P_s}(s)$ denotes the probability that $s \in S$ initiates a rumor spreading process. For normalization, we have

$$\sum_{i=1}^{k} \mathbf{P_s}(s_i) = 1. \tag{1}$$

In the following, we call the nodes in S as *suspect nodes* and assume a uniform *a priori* distribution $\mathbf{P_s}$ over the nodes in S, i.e., $\mathbf{P_s}(s) = 1/k$ for any $s \in S$.

The rumor spreading process unfolds as follows. Initially, only a single node $s^* \in S$ possesses a rumor to spread over the network G, and thus is termed infected. An infected node may infect its neighbors, independent of all other nodes. Let τ_{ij} be the time it takes for node j to receive the rumor from its neighbor i after i has the rumor, where $(i, j) \in E$. In this model, we suppose that $\{\tau_{ij}, (i, j) \in E\}$ are mutually independent and all exponentially distributed with rate λ. Without loss of generality, we assume $\lambda = 1$.

B. Rumor Source Estimator: Maximum a Posteriori (MAP)

(1) MAP estimator

Suppose that a rumor originates from a node $s^* \in S$, and we get to observe the network G at some point in time and find a snapshot of n infected nodes carrying the rumor, which are collectively denoted by G_n. Due to the SI model, G_n must form a connected subgraph of G and contain at least a node in S, that is the rumor source. Our goal is to construct an

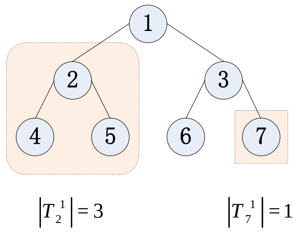

Fig. 1. Illustration of subtree T_u^v.

estimator to identify a node $\hat{s}$ as the estimate of the rumor source s^*.

Now, conditioned on G_n, the source node has a uniform distribution over the nodes in $S \bigcap G_n$ to initiate the infection. Utilizing the Bayes rule, the maximum *a posteriori* (MAP) estimator of s^* that maximizes the correct detection probability is

$$\begin{aligned} \hat{s} &\in \underset{s\in\{S \bigcap G_n\}}{\arg\max}\, \mathbf{P_G}\left(s|G_n\right) \\ &= \underset{s\in\{S \bigcap G_n\}}{\arg\max}\, \frac{\mathbf{P_G}\left(G_n|s\right)\mathbf{P_s}(s)}{\mathbf{P_G}\left(G_n\right)} \\ &= \underset{s\in\{S \bigcap G_n\}}{\arg\max}\, \mathbf{P_G}\left(G_n|s\right), \end{aligned} \tag{2}$$

where $\mathbf{P_G}(G_n|s)$ is the probability of observing G_n, assuming s to be the rumor source. Note that a key difference from the model in [1] is that here we restrict the rumor source to be within the *a priori* suspect set $S \subseteq V$. Naturally, we would first evaluate $\mathbf{P_G}\left(G_n|s\right)$ for all $s \in \{S \bigcap G_n\}$ and then pick the one with the maximal value as the estimate $\hat{s}$.

(2) Optimal MAP estimator for regular trees

In general, the evaluation of $\mathbf{P_G}\left(G_n|s\right)$ may be computationally prohibitive since it is related to counting the number of linear extensions of a partially ordered set [1], [22]. Therefore, we leverage the concept of rumor centrality, first developed in [1], to design our estimator using only $O(n)$ computations.

Similar to the analysis in [1], the optimal MAP estimator for a regular tree can be derived as

$$\hat{s} \in \underset{s\in\{S \bigcap G_n\}}{\arg\max}\, \mathbf{P_G}\left(G_n|s\right) = \underset{s\in\{S \bigcap G_n\}}{\arg\max}\, R\left(s, G_n\right), \tag{3}$$

where $R(s, G_n)$ is the rumor centrality of node s in G_n and can be computed by

$$R(s, G_n) = n! \prod_{u \in G_n} \frac{1}{|T_u^s|}, \tag{4}$$

where T_u^s is the subtree rooted at node u with node s as the source in G_n and $\left|T_u^s\right|$ is the number of nodes in T_u^s; e.g., see Fig. 1.

(3) Approximate estimator for general trees and graphs

For general trees, the estimator (3) with the rumor centrality is also applicable though heuristic. So, the approximate estimator becomes

$$\hat{s} \in \underset{s\in\{S \bigcap G_n\}}{\arg\max}\, R\left(s, G_n\right). \tag{5}$$

Besides, a message-passing algorithm is proposed in [1] to compute rumor centralities for all nodes in a general tree G_n with n nodes, using only $O(n)$ computations.

For general graphs without a tree structure, we can not directly use the estimator (3) with the rumor centrality. With the intuition that the rumor generally travels from the source to each infected node along a minimum-distance path [1], [20], we construct a breadth-first search (BFS) tree as the diffusion tree at first. So, the approximate estimator becomes

$$\hat{s} \in \underset{s\in\{S \bigcap G_n\}}{\arg\max}\, R\left(s, T_{\mathrm{bfs}}(s)\right), \tag{6}$$

where $T_{\mathrm{bfs}}(s)$ is the BFS tree that begins with node s as the source in a general graph G_n with n nodes. In total, at most $O(n^3)$ computations are used.

C. Local Rumor Center on General Trees

The rumor centrality is an important quantity of graph score, reflecting many interesting properties of network structure, such as explicit succinct representation, definite relation between neighboring nodes, equivalence to distance centrality on general trees; see the details in [1].

Next, we develop a concept of local rumor center related to the rumor centrality, which enables efficient implementation of the estimator (3) and furthermore will be instrumental for our subsequent analysis. A useful fact from [1] is that the following relationship holds for any two neighboring nodes u and s in a tree G_n:

$$R(u, G_n) = R(s, G_n)\frac{|T_u^s|}{n - |T_u^s|}. \tag{7}$$

Now, consider a node s with a neighbor set $N(s)$ and a sub-neighborhood $N_l(s) \subseteq N(s)$. If $R(s, G_n) \geq R(u, G_n)$ for all $u \in N_l(s)$, then s is called the local rumor center with respect to (w.r.t.) the sub-neighborhood $N_l(s)$ of G_n. For local rumor center, we have the following proposition.

Proposition 1. *i) Given a tree G_n of n node, if node s^* is the local rumor center w.r.t. a sub-neighborhood $N_l(s^*) \subset G_n$, then for any $u \in N_l(s^*)$, we have $\left|T_u^{s^*}\right| \leq n/2$; and for any $u' \in T_u^{s^*} \setminus \{u\}$, we have $R(u', G_n) < R(s^*, G_n)$.*
ii) If there is a node s^* such that $\left|T_u^{s^*}\right| \leq n/2$ for all $u \in N_l(s^*)$, then s^* is a local rumor center w.r.t. the sub-neighborhood $N_l(s^*) \subset G_n$.
iii) Furthermore, if node s^* is the local rumor center w.r.t. a sub-neighborhood $N_l(s^*) \subset G_n$, then there is at most a node $u \in N_l(s^*)$ such that $R(u, G_n) = R(s^*, G_n)$, which holds if and only if $\left|T_u^{s^*}\right| = n/2$.

Remark 1: In fact, the local rumor center is a generalization of the rumor center in [1], which is the node with the maximal rumor centrality in G_n. When $N_l(s) = N(s)$, the two are the same. However, the rumor center may belong to the set $G_n \setminus S$ and thus is not the solution of the estimator (3). Besides, notice that we can find at most two local rumor centers from a connected suspect set S w.r.t. the sub-neighborhood formed by S, given a snapshot infection G_n.

Proof of Proposition 1. First of all, consider $u \in N_l(s^*)$, by (7) we have

$$\frac{R(u,G_n)}{R(s^*,G_n)} = \frac{\left|T_u^{s^*}\right|}{n-\left|T_u^{s^*}\right|}. \tag{8}$$

Since $R(u,G_n) \le R(s^*,G_n)$, thus $\left|T_u^{s^*}\right| \le n/2$.

Then consider $u' \in T_u^{s^*} \setminus \{u\}$, and let $\mathcal{P}(u,u')$ be the set of nodes along the shortest path from u and u', not including u. Repeatedly using (7), we have

$$\frac{R(u',G_n)}{R(u,G_n)} = \prod_{v\in\mathcal{P}(u,u')} \frac{\left|T_v^{s^*}\right|}{n-\left|T_v^{s^*}\right|}. \tag{9}$$

As we have already known that $\left|T_u^{s^*}\right| \le n/2$, thus $\left|T_v^{s^*}\right| < \left|T_u^{s^*}\right| \le n/2$, i.e. $\left|T_v^{s^*}\right|/(n-\left|T_v^{s^*}\right|) < 1$ for all $v \in \mathcal{P}(u,u')$. Therefore, $R(u',G_n) < R(u,G_n) \le R(s^*,G_n)$. This proves the first part of Proposition 1.

Now for the second part of Proposition 1, if $\left|T_u^{s^*}\right| \le n/2$, i.e. $\left|T_u^{s^*}\right|/(n-\left|T_u^{s^*}\right|) \le 1$ for all $u \in N_l(s^*)$, then by (8) we see that $R(u,G_n) \le R(s^*,G_n)$. Therefore, s^* is the local rumor center w.r.t. its sub-neighborhood $N_l(s^*) \subset G_n$.

As for the third part of Proposition 1, if there is a node $u \in N_l(s^*)$ such that $R(u,G_n) = R(s^*,G_n)$, then by (8) we see that $\left|T_u^{s^*}\right| = n/2$; the deducing process holds backward. Besides, there can be at most a subtree $T_u^{s^*}$ ($u \in N_l(s^*)$) such that $\left|T_u^{s^*}\right| = n/2$, since the total number of nodes in G_n is n and any two subtrees with s^* as the source are disjoint. As a result, if node s^* is the local rumor center w.r.t. a sub-neighborhood $N_l(s^*) \subset G_n$ and there has already been a node $u^* \in N_l(s^*)$ such that $R(u^*,G_n) = R(s^*,G_n)$, then for all $u \in N_l(s^*) \setminus \{u^*\}$ we have $\left|T_u^{s^*}\right| < n/2$, i.e. $\left|T_u^{s^*}\right|/(n-\left|T_u^{s^*}\right|) < 1$. Again, by (8) we have $R(u,G_n) < R(s^*,G_n)$. □

III. Detection Probability on Regular Trees

In this section, we analyze the performance of the MAP rumor source estimator for regular tree-type networks. Three interesting scenarios are analyzed: First, when the suspect set contains all nodes in the network, the best detection probability is 0.307 without any *a priori* knowledge. Second, when the suspect set forms a small connected subgraph of the network, reliable detection can be achieved. Third, when the suspect set contains only two nodes, the detection probability increases with their separation distance. We show that the rumor spreading process on regular trees is equivalent to the ball drawing process of the Pólya's urn model, and utilize its well-known distribution to establish the performance of the MAP estimator. In the following, we start by introducing our main results and the preliminaries on the Pólya's urn model.

A. Main Results

We focus on the correct detection probability $\mathbf{P_c}(n)$, the probability of the estimator to correctly identify the rumor source from the suspect set S, upon observing G_n of n infected nodes across the network $G = (V,E)$. For a regular tree with node degree δ, we can use numerical computation to calculate $\mathbf{P_c}(n)$ as n is small. Besides, we have the following three theorems for $\mathbf{P_c}(n)$.

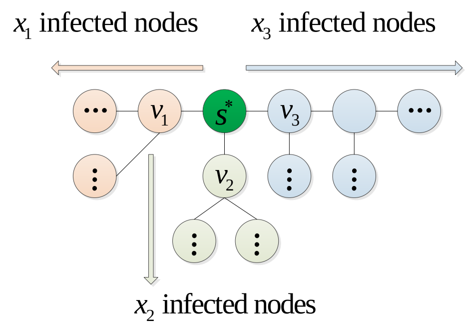

Fig. 2. Illustration of $X_1 = x_1$, $X_2 = x_2$, $X_3 = x_3$ infected nodes in subtrees $T_{v_1}^{s^*}$, $T_{v_2}^{s^*}$ and $T_{v_3}^{s^*}$ with s^* as the source across a regular tree G with node degree $\delta = 3$, respectively.

(1) Suspecting all nodes

In the extreme case of $S = V$, every infected node in an observation G_n has the potential to be the rumor source, i.e. without any *a priori* knowledge. Assuming a node s^* to be the rumor source, we could observe $X_j = x_j$ ($1 \le j \le \delta$) infected nodes in each subtree $T_u^{s^*}$ rooted at node v_j with node s^* as the source in G_n, where X_j is a random variable; e.g., see Fig. 2. Then, we will establish the following theorem.

Theorem 2. *Suppose $S = V$, i.e., every infected node is a suspect node, then:*
i) When $\delta = 2$ (line graph),

$$\mathbf{P_c}(n) = \frac{1}{2^{n-1}}\binom{n-1}{\lfloor (n-1)/2 \rfloor}, \tag{10}$$

and $\mathbf{P_c}(n) = O(1/\sqrt{n})$ with sufficiently large n.
ii) When $\delta = 3$,

$$\mathbf{P_c}(n) = \frac{1}{4} + \frac{3}{4}\frac{1}{2\lfloor n/2 \rfloor + 1}, \tag{11}$$

and $\mathbf{P_c}(n) = 0.25 + O(1/n)$ with sufficiently large n.
iii) When $\delta > 3$,

$$\lim_{n\to\infty} \mathbf{P_c}(n) = \phi_1(\delta) := 1 - \delta\left(1 - I_{1/2}\left(\frac{1}{\delta-2}, \frac{\delta-1}{\delta-2}\right)\right), \tag{12}$$

where $I_x(\alpha,\beta)$ is the regularized incomplete Beta function with parameters α and β; and $\phi_1(\delta) = 1 - \ln 2 \approx 0.307$ as δ grows sufficiently large.

Remark 2: With $S = V$, the estimator (3) in fact reduces to the ML estimator established in [1], [2].[1] The asymptotic parts of Theorem 1 have been established in [1], [2] via a rather different approach from ours, explicitly relying upon the exponential distribution of the infection time. We see that without any *a priori* knowledge, the estimator achieves nontrivial detection probability if the network is not linear, but the asymptotic detection probability is upper bounded by 0.307.

(2) Connected suspects

[1]Strictly speaking, our setup of uniform *a priori* distribution of the rumor source over S is not well defined when $S = V$ since it is a countably infinite set. Our remedy is that we consider ML estimation for Theorem 2 thus returning to the situation in [1], [2], and consider MAP estimation for the other two cases.

In the case where $S = \{s_1, s_2, \dots, v_k\}$ with cardinality k forms a connected subgraph of the network G; e.g., see Fig. 3. With this *a priori* knowledge, we will establish the following theorem.

Theorem 3. *Suppose that S forms a connected subgraph of G, and assume $|S| = k \ll n$, then:*
i) When $\delta = 2$ (line graph),

$$\mathbf{P_c}(n) = \frac{1}{k}\left(1 + \frac{k-1}{2^{n-1}}\binom{n-1}{\lfloor (n-1)/2 \rfloor}\right), \tag{13}$$

and $\mathbf{P_c}(n) = 1/k + O(1/\sqrt{n})$ with sufficiently large n.
ii) When $\delta = 3$,

$$\mathbf{P_c}(n) = \frac{k+1}{2k} + \frac{k-1}{k}\frac{1}{4\lfloor n/2 \rfloor + 2}, \tag{14}$$

and $\mathbf{P_c}(n) = (k+1)/(2k) + O(1/n)$ with sufficiently large n.
iii) When $\delta > 3$,

$$\lim_{n\to\infty} \mathbf{P_c}(n) = \phi_2(\delta, k) := 1 - \frac{2k-2}{k}\left(1 - I_{1/2}\left(\frac{1}{\delta-2}, \frac{\delta-1}{\delta-2}\right)\right), \tag{15}$$

and $\phi_2(\delta, k) = 1$ as δ grows sufficiently large.

Remark 3: For line graphs, $\mathbf{P_c}(n)$ can barely exceed the *a priori* probability $1/k$. When $\delta \geq 3$, $\lim_{n\to\infty} \mathbf{P_c}(n) \geq 1/k + (k-1)/(2k) > 1/k$ for all $k \ll n$. Furthermore, the MAP estimator achieves reliable detection as δ grows sufficiently large. Therefore, we see that the performance of the MAP estimator is significantly improved and reliable detection can be achieved, only given the *a priori* knowledge on a small connected subgraph composed by the suspect nodes.

(3) Two suspects

In the case where S contains only two suspect nodes s_1 and s_2, denote by d their shortest path distance on the network G; e.g., see Fig. 4. Since we can ensure correct estimation of the source if $d \geq n$, we assume $d < n$. With this *a priori* knowledge, we have the following theorem.

Theorem 4. *Suppose S only contains two suspect nodes, and denote by d their shortest path distance ($d < n$), then:*
i) When $\delta = 2$ (line graph),

$$\mathbf{P_c}(n) = \begin{cases} \frac{1}{2} - \frac{1}{2^n}\sum_{z_1=(n-d-1)/2}^{(n+d+1)/2}\binom{n-1}{z_1}, & (n-d) \text{ is odd;} \\ \frac{1}{2} - \frac{1}{2^n}\sum_{z_1=(n-d)/2}^{(n+d-2)/2}\binom{n-1}{z_1}, & (n-d) \text{ is even.} \end{cases} \tag{16}$$

ii) When $\delta = 3$,

$$\lim_{n\to\infty} \mathbf{P_c}(n) \begin{cases} = 0.75, & d = 1, \\ \approx 0.886, & d = 2. \end{cases} \tag{17}$$

iii) When $\delta > 3$,

$$\lim_{n\to\infty} \mathbf{P_c}(n) = \phi_3(\delta) := I_{1/2}\left(\frac{1}{\delta-2}, \frac{\delta-1}{\delta-2}\right), \; d = 1, \tag{18}$$

and $\phi_3(\delta) = 1$ as δ grows sufficiently large.
iv) In general $\mathbf{P_c}(n)$ increases with d.

Remark 4: It is harder to correctly identify the rumor source if the two suspect nodes are closer, i.e. with stronger social strength, to each other; and vice versa. Besides, the *a priori* probability of $1/2$ can be also significantly exceeded when $\delta \geq 3$. Furthermore, the MAP estimator also achieves reliable detection whenever δ grows sufficiently large.

B. Equivalence to the Pólya's Urn Model

(1) Preliminaries

Before the argument of the theorems above, we introduce the preliminaries on the Pólya's urn model, then we will show that the rumor spreading process on regular trees is equivalent to the ball drawing process of the Pólya's urn model, a fact first noticed and utilized in [2]. This allows the exact detection probability to be explicitly established.

Pólya's urn model ([21], Chapter 4): initially, the urn contains b_j balls of color C_j ($1 \leq j \leq \delta$); at each uniform drawing of a single ball, the ball is returned together with s balls of the same color; after n draws, the number X_j is the number of times that the balls of color C_j are drawn. Then

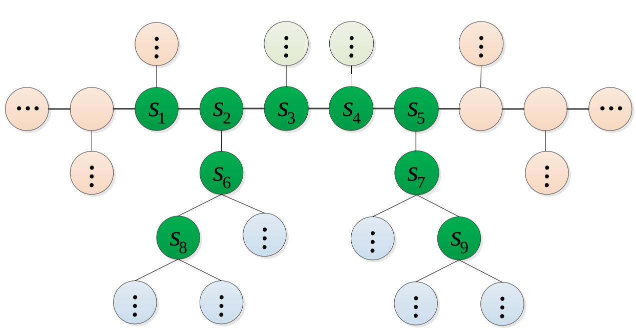

Fig. 3. Illustration of a suspect set $S = \{s_1, s_2, \dots, s_9\}$ with multiple connected suspect nodes across a regular tree G with node degree $\delta = 3$. The nodes in S form a connected subgraph of G.

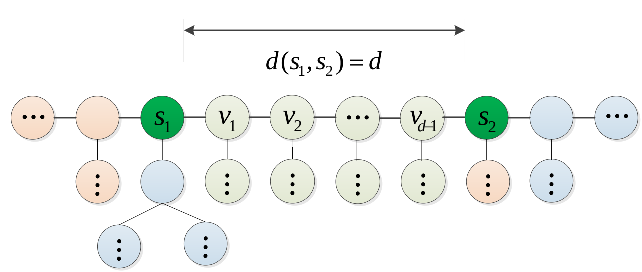

Fig. 4. Illustration of two suspect nodes with distance d on a regular tree G with node degree $\delta = 3$. There are d nodes from s_1 to s_2 except s_1, which are sequentially notated as $v_1, v_2, \dots, v_{d-1}, v_d = s_2$.

the joint distribution of $\{X_j, 1 \le j \le \delta\}$ can be derived as

$$\mathbf{P_G}\left[\bigcap_{j=1}^{\delta}(X_j = x_j)\right] = \frac{n!}{x_1!x_2!\cdots x_\delta!}\frac{\prod_{j=1}^{\delta} b_j(b_j+s)\cdots(b_j+(x_j-1)s)}{b(b+s)\cdots(b+(n-1)s)}, \quad (19)$$

where $b = \sum_{j=1}^{\delta} b_j$ and $\sum_{j=1}^{\delta} x_j = n$. In fact, it is a $\delta - 1$ dimensional distribution.

As $n \to \infty$, the limiting joint distribution of the ratios $\{X_j/n, 1 \le j \le \delta\}$ converges to the Dirichlet distribution with a density function given by

$$\lim_{n\to\infty} \mathbf{P_G}\left[\bigcap_{j=1}^{\delta}\left(\frac{X_j}{n} = y_j\right)\right] = \frac{\Gamma(\alpha)}{\prod_{j=1}^{\delta}\Gamma(\alpha_j)}\prod_{j=1}^{\delta} y_j^{\alpha_j - 1}, \quad (20)$$

where $\alpha_j = b_j/s$, $\alpha = \sum_{j=1}^{\delta}\alpha_j$ and $\sum_{j=1}^{\delta} y_j = 1$. Here, $\Gamma(\alpha)$ is the gamma function with parameter α.

Besides, the marginal distribution of X_1 is

$$\mathbf{P_G}(X_1 = x_1) = \frac{n!}{x_1'!x_2'!}\frac{\prod_{j=1}^{2} b_j'(b_j'+s)\cdots(b_j'+(x_j'-1)s)}{b(b+s)\cdots(b+(n-1)s)}, \quad (21)$$

where $b_1' = b_1$, $b_2' = b - b_1$, $x_1' = x_1$ and $x_2' = n - x_1$. In fact, it can be seen as a special case of the Pólya's urn model with two colors.

As $n \to \infty$, the limiting marginal distribution of the ratio X_1/n converges to the Beta distribution with a density function given by

$$\lim_{n\to\infty} \mathbf{P_G}\left(\frac{X_1}{n} = y_1\right) = \frac{\Gamma(\alpha_1' + \alpha_2')}{\Gamma(\alpha_1')\Gamma(\alpha_2')} y_1^{\alpha_1'-1}(1-y_1)^{\alpha_2'-1}, \quad (22)$$

where $\alpha_1' = \alpha_1$ and $\alpha_2' = \alpha - \alpha_1$. The cumulative distribution function of the Beta distribution is

$$I_x(\alpha_1', \alpha_2') := \frac{\Gamma(\alpha_1' + \alpha_2')}{\Gamma(\alpha_1')\Gamma(\alpha_2')}\int_0^x y^{\alpha_1'-1}(1-y)^{\alpha_2'-1}dy, \quad (23)$$

for all $x \in [0, 1]$. $I_x(\alpha_1', \alpha_2')$ is called as the regularized incomplete Beta function with parameters α_1' and α_2'.

In particular, we are interested in $I_{1/2}(\alpha_1', \alpha_2')$ with parameters $\alpha_1' = 1/(\delta - 2)$ and $\alpha_2' = (\delta - 1)/(\delta - 2)$, where δ is the node degree of a regular tree and $\delta \ge 3$; e.g., see Fig. 5.

(2) Equivalence to the Pólya's urn model

Next, we show that the rumor spreading process on regular trees is equivalent to the ball drawing process of the Pólya's urn model, whose well-known distributions will be used in our subsequent analysis for Theorem 2 and Theorem 3.

For a rumor source s^* with δ neighboring nodes $v_1, \ldots, v_\delta$, let $T_{v_j}^{s^*}$ $(1 \le j \le \delta)$ be the subtree rooted at node v_j with node s^* as the source in G_n, and define a random variable X_j as the number of nodes in $T_{v_j}^{s^*}$; e.g., see Fig. 2. In the rumor spreading process, nodes in G_n are infected sequentially, and thus we have the following: initially, s^* has one neighbor in each subtree $T_{v_j}^{s^*}$ $(1 \le j \le \delta)$ that can be potentially infected; after one of those nodes is infected, it introduces $\delta - 1$ new potentially susceptible nodes; finally, $n - 1$ nodes are infected besides s^*. Due to the memoryless and i.i.d. properties of exponentially distributed infection times $\{\tau_{ij}, (i, j) \in E\}$, in each step, the infected node is uniformly selected from the susceptible nodes.

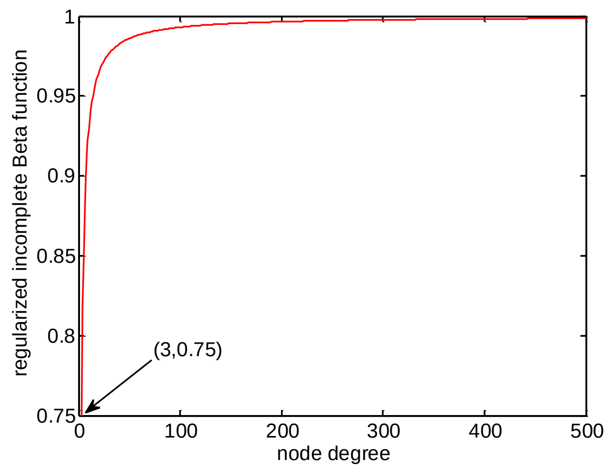

Fig. 5. The regularized incomplete Beta function $I_{1/2}(1/(\delta-2), (\delta-1)/(\delta-2))$ vs. node degree δ.

Now, the resulting infection G_n with X_j nodes in $T_{v_j}^{s^*}$ $(1 \le j \le \delta)$ can be constructed in an equivalent way by the Pólya's urn model [21, Chapter 4]: initially, the urn has one ball for each color C_j; at each uniform drawing of a single ball, the ball is returned together with $s = \delta - 2$ additional balls of the same color; after $n - 1$ draws, X_j is the number of times that the balls of color C_j are drawn.

Therefore, in the rumor spreading process, if we assume s^* to be the rumor source with δ neighbors $v_1, v_2, \ldots, v_\delta$ and observe n infected nodes G_n with X_j nodes in each subtree $T_{v_j}^{s^*}$ $(1 \le j \le \delta)$, then by (19) the joint distribution of $\{X_j, 1 \le j \le \delta\}$ can be derived as

$$\mathbf{P_G}\left[\bigcap_{j=1}^{\delta}(X_j = x_j)\right] = \frac{(n-1)!}{x_1!x_2!\cdots x_\delta!}\frac{\prod_{j=1}^{\delta} 1(1+s)\cdots(1+(x_j-1)s)}{\delta(\delta+s)\cdots(\delta+(n-2)s)}, \quad (24)$$

where $\sum_{j=1}^{\delta} x_j = n - 1$.

We are also interested in the limiting marginal distribution of the ratio X_1/n as $n \to \infty$. By (22), we have

$$\lim_{n\to\infty} \mathbf{P_G}\left(\frac{X_1}{n} = y\right) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1}, \quad (25)$$

where $\alpha = 1/(\delta-2)$, $\beta = (\delta-1)/(\delta-2)$. By the way, this limit distribution is also used in [2] to derive the asymptotic correct detection probability in the case $S = V$ with $\delta > 3$.

(3) Markov concatenation of the Pólya's urn model

For Theorem 4, we use the Markov concatenation of the Pólya's urn model to derive the probability distribution of the infection observed in the rumor spreading process, focusing on the nodes along the path from s_1 and s_2.

Assume s_1 to be the rumor source s^*, and let $\mathcal{P} = \{v_0 = s_1, v_1, \cdots, v_d = s_2\}$ be the shortest path from s_1 to s_2; e.g., see Fig. 4. Here, let a random variable Z_h $(1 \le h \le d)$ be

the number of nodes in subtree $T_{v_h}^{s^*}$ rooted at node v_h with node s^* as the source in G_n. It is clear that $Z_h \geq Z_{h+1}+1$ if $Z_h > 0$ for all $1 \leq h \leq d-1$. In the proof of Theorem 4, we focus on the error detection probability $\mathbf{P_e}(n) = 1 - \mathbf{P_c}(n)$. Therefore, without loss of generality we assume $Z_h > 0$ for all $1 \leq h \leq d$. Here, we can also construct the random variables Z_h ($1 \leq h \leq d$) equivalently with the concatenation use of the Pólya's urn model.

For random variable Z_1, we construct it by the Pólya's urn model as follows: initially, the urn has $b_b^1 = 1$ black ball and $b_w^1 = \delta - 1$ white balls; at each uniform drawing of a single ball, the ball is returned together with $s = \delta - 2$ additional balls of the same color; after $n-1$ draws, the number Z_1 is the number of times that the balls of black color are drawn. Therefore, by (19) the distribution of Z_1 is

$$\mathbf{P_G}\left[Z_1 = z_1\right] = \binom{n-1}{z_1}\frac{b_b^1\cdots(b_b^1+(z_1-1)s)b_w^1\cdots(b_w^1+(n-z_1-2)s)}{\delta(\delta+s)\cdots(\delta+(n-2)s)}. \quad (26)$$

For random variable Z_h ($2 \leq h \leq d$) conditioned on $Z_{h-1} = z_{h-1}$, we construct it by the Pólya's urn model as follows: initially, the urn has $b_b^h = 1$ black ball and $b_w^h = \delta - 2$ white balls; at each uniform drawing of a single ball, the ball is returned together with $s = \delta - 2$ additional balls of the same color; after $z_{h-1} - 1$ draws, the number Z_h is the number of times that the balls of black color are drawn. Again, by (19) the distribution of Z_h is

$$\mathbf{P_G}\left[Z_h = z_h | Z_{h-1} = z_{h-1}\right] = \binom{z_{h-1}-1}{z_h}\frac{b_b^h\cdots(b_b^h+(z_h-1)s)b_w^h\cdots(b_w^h+(z_{h-1}-z_h-2)s)}{(\delta-1)(\delta-1+s)\cdots(\delta-1+(z_{h-1}-2)s)}. \quad (27)$$

Since Z_h only depends on Z_{h-1} for all $2 \leq h \leq d$, thus $Z_1, Z_2, \ldots, Z_d$ form a Markov chain. Therefore, by the Markov chain rule the joint distribution of $\{Z_h, 1 \leq h \leq d\}$ is

$$\mathbf{P_G}\left[\bigcap_{h=1}^{d}(Z_h = z_h)\right] = \mathbf{P_G}\left[Z_1 = z_1\right]\prod_{h=2}^{d}\mathbf{P_G}\left[Z_h = z_h | Z_{h-1} = z_{h-1}\right]. \quad (28)$$

C. Proof of Theorem 2: Suspecting all Nodes

In the case of $S = V$, we only need to consider an arbitrary node $s^* \in G$ as the rumor source by symmetry. For a source s^* with m ($m \leq \delta$) neighboring nodes $N_l(s^*) = \{s_1^*, \ldots, s_m^*\} \subset S$, let a random variable X_j be the number of nodes in each subtree $T_{s_j^*}^{s^*}$ ($1 \leq j \leq m$) of G_n. Then, we have the following lemma for the argument of Theorem 2 ($m = \delta$) and Theorem 3 ($m \leq \delta$); see its proof in Appendix-A.

Lemma 5. *To correctly identify source s^* with m neighboring suspect nodes as the estimate $\hat{s}$, we have*

$$\begin{cases} p_1 := \mathbf{P_c}\left(\hat{s} = s^* \middle| \max\{x_j, 1 \leq j \leq m\} < n/2\right) = 1, \\ p_{1/2} := \mathbf{P_c}\left(\hat{s} = s^* \middle| \max\{x_j, 1 \leq j \leq m\} = n/2\right) = \dfrac{1}{2}, \\ p_0 := \mathbf{P_c}\left(\hat{s} = s^* \middle| \max\{x_j, 1 \leq j \leq m\} > n/2\right) = 0. \end{cases} \quad (29)$$

Remark 5: Lemma 5 is deduced from Proposition 1. In order to prove Theorem 2 (and Theorem 3), we should find the conditions, under which s^* is the local rumor center w.r.t. $N_l(s^*)$ of G_n, such that the estimator (3) can correctly identify s^* as the source.

Here, $S = V$, thus source s^* has $m = \delta$ neighboring suspect nodes. By Lemma 5, we can write $\mathbf{P_c}(n)$ as

$$\begin{aligned}\mathbf{P_c}(n) &= p_{1/2}\cdot\sum_{\max\{x_j,1\leq j\leq\delta\}=n/2}\mathbf{P_G}\left[\bigcap_{j=1}^{\delta}(X_j = x_j)\right] \\ &\quad + p_1\cdot\sum_{\max\{x_j,1\leq j\leq\delta\}<n/2}\mathbf{P_G}\left[\bigcap_{j=1}^{\delta}(X_j = x_j)\right], \end{aligned} \quad (30)$$

where $\mathbf{P_G}\left[\bigcap_{j=1}^{\delta}(X_j = x_j)\right]$ is the probability of observing n infected nodes G_n with $X_j = x_j$ nodes in subtree $T_{v_j}^{s^*}$ for all $1 \leq j \leq \delta$ on a regular tree with node degree δ; see its closed-form expression in (24). Notice that $T_{v_j}^{s^*} = T_{s_j^*}^{s^*}$ for all $1 \leq j \leq \delta$.

As n is small, we can make use of numerical computation to calculate the correct detection probability $\mathbf{P_c}(n)$ with (30). In the following, we will establish the argument of Theorem 2 under three situations when the node degree $\delta = 2$, $\delta = 3$ and $\delta > 3$, respectively.

(1) Detection probability when $\delta = 2$

Proof of Theorem 2-i. When $\delta = 2$, the distribution in (24) can be written as

$$\mathbf{P_G}\left[\bigcap_{j=1}^{2}(X_j = x_j)\right] = \frac{(n-1)!}{x_1!x_2!}\frac{1}{2^{n-1}}. \quad (31)$$

By (30), the correct detection probability is

$$\begin{aligned}\mathbf{P_c}(n) &= p_{1/2}\cdot\sum_{\max\{x_1,x_2\}=n/2}\mathbf{P_G}\left[\bigcap_{j=1}^{2}(X_j = x_j)\right] \\ &\quad + p_1\cdot\sum_{\max\{x_1,x_2\}<n/2}\mathbf{P_G}\left[\bigcap_{j=1}^{2}(X_j = x_j)\right], \\ &= \frac{1}{2^{n-1}}\binom{n-1}{\lfloor(n-1)/2\rfloor}. \end{aligned} \quad (32)$$

Above, the detailed deduction is in Appendix-B.

As $n \to \infty$, by the Stirling's formula we have

$$\begin{aligned}\mathbf{P_c}(n) &\approx \frac{1}{2^n}\cdot\frac{n!}{[(n/2)!]^2} \\ &\approx \frac{1}{2^n}\cdot\frac{\sqrt{2\pi n}\cdot\left(\frac{n}{e}\right)^n}{\left[\sqrt{\pi n}\cdot\left(\frac{n}{2e}\right)^{n/2}\right]^2} \\ &= \sqrt{\frac{2}{\pi n}} \\ &= O\left(\frac{1}{\sqrt{n}}\right). \end{aligned} \quad (33)$$

□

(2) Detection probability when $\delta = 3$

Proof of Theorem 2-ii. When $\delta = 3$, the distribution in (24) can be written as

$$\mathbf{P_G}\left[\bigcap_{j=1}^{3}(X_j = x_j)\right] = \frac{2}{n(n+1)}. \tag{34}$$

By (30), the correct detection probability is

$$\begin{aligned}\mathbf{P_c}(n) &= p_{1/2}\cdot \sum_{\max\{x_j,1\le j\le 3\}=n/2} \mathbf{P_G}\left[\bigcap_{j=1}^{3}(X_j = x_j)\right] \\ &\quad + p_1 \cdot \sum_{\max\{x_j,1\le j\le 3\}<n/2} \mathbf{P_G}\left[\bigcap_{j=1}^{3}(X_j = x_j)\right] \\ &= \frac{1}{4} + \frac{3}{4}\frac{1}{2\lfloor n/2\rfloor + 1}. \end{aligned} \tag{35}$$

Above, the detailed deduction is in Appendix-C. □

(3) Detection probability when $\delta > 3$

As n is small, we can make use of numerical computation to calculate the correct detection probability with (30). As $n \to \infty$, we derive the asymptomatic correct detection probability in a similar way as that in [2].

Before the argument of Theorem 2-iii, we present the following lemma, which will be also used in the argument of Theorem 3-iii; see its proof in Appendix-D. For a source s^* with m ($m \le \delta$) neighboring nodes $N_l(s^*) = \{s_1^*, \ldots, s_m^*\}$ in S, let a random variable X_j be the number of nodes in each subtree $T_{s_j^*}^{s^*}$ ($1 \le j \le m$) of G_n.

Lemma 6. *Define $E_j = \{X_j < n/2\}$ and $F_j = \{X_j \le n/2\}$, $1 \le j \le m$. To correctly identify source s^* with m neighboring suspect nodes, we have*

$$\begin{cases}\mathbf{P_c}(n|s^*) \ge 1 - m\mathbf{P_G}(E_1^c) \\ \mathbf{P_c}(n|s^*) \le 1 - m\mathbf{P_G}(F_1^c)\end{cases} \tag{36}$$

where $\mathbf{P_G}(E_1^c)$ and $\mathbf{P_G}(F_1^c)$ are the probabilities that the complements of events E_1 and F_1 occur, respectively.

Remark 6: Lemma 6 is deduced from Proposition 1, and is a generalization of the statement claimed in [2, Section 4.1.2]. In order to prove Theorem 2-iii (and Theorem 3-iii), we should show that the lower and upper bounds asymptotically coincide.

Proof of Theorem 2-iii. Since $E_1 = \{X_1 < n/2\}$ and $F_1 = \{X_1 \le n/2\}$, i.e., $E_1 = \{X_1/n < 1/2\}$ and $F_1 = \{X_1/n \le 1/2\}$, we consider the cumulative distribution function of the ratio X_1/n as $n \to \infty$; see the expression in (25). Then,

$$\begin{aligned}\lim_{n\to\infty}\mathbf{P_G}(E_1) &= \lim_{n\to\infty}\mathbf{P_G}(F_1) \\ &= \int_{y=0}^{1/2}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}y^{\alpha-1}(1-y)^{\beta-1}dy \\ &= I_{1/2}\left(\frac{1}{\delta-2}, \frac{\delta-1}{\delta-2}\right),\end{aligned} \tag{37}$$

where $I_x(\alpha,\beta)$ is the regularized incomplete Beta function with parameters $\alpha = 1/\delta$ and $\beta = (\delta-1)/(\delta-2)$.

Since $S = V$, the source s^* has $m = \delta$ neighboring suspect nodes. By Lemma 6, the correct detection probability is

$$\mathbf{P_c}(n) = 1 - \delta\left(1 - I_{1/2}\left(\frac{1}{\delta-2}, \frac{\delta-1}{\delta-2}\right)\right), \tag{38}$$

as $n \to \infty$.

Besides, we see that the asymptotic value $\lim_{n\to\infty}\mathbf{P_c}(n) > 0.25$ if the node degree $\delta > 3$. Furthermore, it approaches $1 - \ln 2 \approx 0.307$ as $\delta \to \infty$. □

D. Proof of Theorem 3: Connected Suspects

Now, consider the case where $S = \{s_1, s_2, \ldots, v_k\}$ with cardinality k forms a connected subgraph of the network G. By the Bayes rule and the *a priori* knowledge that $\mathbf{P_s}(s^*) = 1/k$ for any $s^* \in S$, we have

$$\mathbf{P_c}(n) = \sum_{i=1}^{k}\mathbf{P_s}(s_i)\mathbf{P_c}(n|s_i) = \frac{1}{k}\sum_{s^*\in S}\mathbf{P_c}(n|s^*). \tag{39}$$

Therefore, we should first find the correct detection probability $\mathbf{P_c}(n|s^*)$ for each suspect node $s^* \in S$.

Assume that $s^* \in S$ is the rumor source and it has m ($m \le \delta$) neighboring suspect nodes $N_l(s^*) = \{s_1^*, \ldots, s_m^*\} \subset S$. Let a random variable X_j be the number of nodes in each subtree $T_{s_j^*}^{s^*}$ ($1 \le j \le m$) of G_n, then by Lemma 5, we should find the conditions that s^* is the local rumor center w.r.t. $N_l(s^*)$ of G_n and the estimator (3) can correctly identify s^* as the source.

By Lemma 5, we can write $\mathbf{P_c}(n|s^*)$ as

$$\begin{aligned}\mathbf{P_c}(n|s^*) &= p_{1/2}\cdot \sum_{\max\{x_j,1\le j\le m\}=n/2} \mathbf{P_G}\left[\bigcap_{j=1}^{\delta}(X_j = x_j)\right] \\ &\quad + p_1 \cdot \sum_{\max\{x_j,1\le j\le m\}<n/2} \mathbf{P_G}\left[\bigcap_{j=1}^{\delta}(X_j = x_j)\right],\end{aligned} \tag{40}$$

where $\mathbf{P_G}\left[\bigcap_{j=1}^{\delta}(X_j = x_j)\right]$ is the probability of observing n infected nodes G_n with $X_j = x_j$ nodes in subtree $T_{v_j}^{s^*}$ for all $1 \le j \le \delta$ on a regular tree with node degree δ; see its closed-form expression in (24). Notice that $T_{v_j}^{s^*} = T_{s_j^*}^{s^*}$ for all $1 \le j \le m$.

As n is small, we can make use of numerical computation to calculate the correct detection probability $\mathbf{P_c}(n)$ with (39) and (40). In the following, we will establish the argument of Theorem 3 under three situations when the node degree $\delta = 2$, $\delta = 3$ and $\delta > 3$, respectively.

(1) Detection probability when $\delta = 2$

Proof of Theorem 3-i. When $\delta = 2$, the distribution in (24) can be written as

$$\mathbf{P_G}\left[\bigcap_{j=1}^{2}(X_j = x_j)\right] = \frac{(n-1)!}{x_1!x_2!}\frac{1}{2^{n-1}}. \tag{41}$$

By (40), the correct detection probability for a suspect node s^* with m neighbors in the suspect set S is

$$\begin{aligned}\mathbf{P_c}(n|s^*) &= p_{1/2}\cdot \sum_{\max\{x_j,1\le j\le m\}=n/2} \mathbf{P_G}\left[\bigcap_{j=1}^{2}(X_j = x_j)\right] \\ &\quad + p_1 \cdot \sum_{\max\{x_j,1\le j\le m\}<n/2} \mathbf{P_G}\left[\bigcap_{j=1}^{2}(X_j = x_j)\right] \\ &= \begin{cases}\dfrac{1}{2} + \dfrac{1}{2^n}\dbinom{n-1}{\lfloor (n-1)/2\rfloor}, m = 1; \\ \dfrac{1}{2^{n-1}}\dbinom{n-1}{\lfloor (n-1)/2\rfloor}, m = 2.\end{cases}\end{aligned} \tag{42}$$

Above, the detailed deduction is in Appendix-E.

Now for suspect set $S = \{s_1, s_2, \ldots, v_k\}$ with cardinality k, which forms a connected subgraph of a line graph G, i.e., a sub-line graph, we know that only the two suspect nodes at the endpoints of the sub-line graph have one neighboring suspect node and all other suspect nodes have two neighboring suspect nodes. Therefore, by (39) we have

$$\mathbf{P_c}(n) = \frac{1}{k}\left[1 + \frac{k-1}{2^{n-1}}\binom{n-1}{\lfloor (n-1)/2 \rfloor}\right]. \tag{43}$$

As $n \to \infty$, by the same method used in (33) we have

$$\mathbf{P_c}(n) = \frac{1}{k} + \frac{k-1}{k}O(\frac{1}{\sqrt{n}}) = \frac{1}{k} + O(\frac{1}{\sqrt{n}}). \tag{44}$$

□

(2) Detection probability when $\delta = 3$

Proof of Theorem 3-ii. When $\delta = 3$, the distribution in (24) can be written as

$$\mathbf{P_G}\left[\bigcap_{j=1}^{3}(X_j = x_j)\right] = \frac{2}{n(n+1)}. \tag{45}$$

By (40), the correct detection probability for a suspect node s^* with m neighbors in the suspect set S is

$$\begin{aligned}
\mathbf{P_c}(n|s^*) &= p_{1/2} \cdot \sum_{\max\{x_j, 1 \le j \le m\} = n/2} \mathbf{P_G}\left[\bigcap_{j=1}^{3}(X_j = x_j)\right] \\
&\quad + p_1 \cdot \sum_{\max\{x_j, 1 \le j \le m\} < n/2} \mathbf{P_G}\left[\bigcap_{j=1}^{3}(X_j = x_j)\right] \\
&= \begin{cases} \frac{3}{4} + \frac{1}{4}\frac{1}{2\lfloor n/2 \rfloor + 1}, m = 1; \\ \frac{1}{2} + \frac{1}{2}\frac{1}{2\lfloor n/2 \rfloor + 1}, m = 2; \\ \frac{1}{4} + \frac{3}{4}\frac{1}{2\lfloor n/2 \rfloor + 1}, m = 3. \end{cases}
\end{aligned} \tag{46}$$

Above, the detailed deduction is in Appendix-F.

Now for suspect set S with cardinality k, which forms a connected subgraph of a regular tree G with node degree $\delta = 3$, for each suspect node $s^* \in S$ we should first see the number of its neighboring suspect nodes. Notice that, given s^* with m neighboring suspect nodes, $\mathbf{P_c}(n|s^*)$ is one reduced by a same factor, $1/4 - 1/(8\lfloor n/2 \rfloor + 4)$, m times, each of which accounts for one neighboring suspect node of s^* connected by an edge. Since there are $k-1$ edges connecting the k suspect nodes of S in the network G, each edge will account for a reduction of the factor twice. Therefore, by (39) we have

$$\begin{aligned}
\mathbf{P_c}(n) &= 1 - \frac{1}{k}\cdot\left[2(k-1)\cdot\left(\frac{1}{4} - \frac{1}{8\lfloor n/2 \rfloor + 4}\right)\right] \\
&= \frac{k+1}{2k} + \frac{k-1}{k}\frac{1}{4\lfloor n/2 \rfloor + 2}.
\end{aligned} \tag{47}$$

□

(3) Detection probability when $\delta > 3$

As n is small, we can make use of numerical computation to calculate the correct detection probability with (39) and (40). As $n \to \infty$, we can derive the asymptotic correct detection probability with Lemma 6 and an insight analysis into the graph structure.

Based on Lemma 6, we derive $\mathbf{P_c}(n|s^*)$ with its lower and upper bounds, which both asymptotically coincide as $n \to \infty$, in order to prove Theorem 3-iii.

Proof of Theorem 3-iii. When $n \to \infty$, by (37) we have

$$\lim_{n\to\infty}\mathbf{P_G}(E_1) = \lim_{n\to\infty}\mathbf{P_G}(F_1) = I_{1/2}\left(\frac{1}{\delta-2}, \frac{\delta-1}{\delta-2}\right), \tag{48}$$

where $I_x(\alpha, \beta)$ is the regularized incomplete Beta function with parameters α and β.

By Lemma 6, the correct detection probability for a suspect node s^* with m neighbors in the suspect set S is

$$\mathbf{P_c}(n|s^*) = 1 - m\left(1 - I_{1/2}\left(\frac{1}{\delta-2}, \frac{\delta-1}{\delta-2}\right)\right), \tag{49}$$

as $n \to \infty$.

Now for suspect set S with cardinality k, which forms a connected subgraph of a regular tree G with node degree $\delta > 3$, for each suspect node $s^* \in S$ we should first see the number of its neighboring suspect nodes. Notice that, $\lim_{n\to\infty}\mathbf{P_c}(n|s^*)$ is one reduced by a same factor m times, each of which accounts for one neighboring suspect node of s^* connected by an edge. Since there are $k-1$ edges connecting the k suspect nodes of S in the network G, each edge will account for a reduction of the factor twice. Besides, after we reduce the factor by $2(k-1)$ times in total, the limit characteristics in the factor still holds if $k \ll n$. Therefore, by (39) we have

$$\begin{aligned}
\mathbf{P_c}(n) &= 1 - \frac{1}{k}\cdot\left[2(k-1)\cdot\left(1 - I_{1/2}\left(\frac{1}{\delta-2}, \frac{\delta-1}{\delta-2}\right)\right)\right] \\
&= 1 - \frac{2(k-1)}{k} + \frac{2(k-1)}{k}I_{1/2}\left(\frac{1}{\delta-2}, \frac{\delta-1}{\delta-2}\right),
\end{aligned} \tag{50}$$

for all $k \ll n$ and $n \to \infty$.

From Fig. 5, we see that $I_{1/2}\left(\frac{1}{\delta-2}, \frac{\delta-1}{\delta-2}\right) > 0.75$ as the node degree $\delta > 3$. Therefore, we have

$$\mathbf{P_c}(n) > \frac{k+1}{2k}, \tag{51}$$

for all $k \ll n$, $n \to \infty$ and $\delta > 3$.

Besides, $I_{1/2}(0, 1) = 1$ (setting $\delta \to \infty$). Therefore,

$$\mathbf{P_c}(n) = 1, \tag{52}$$

for all $k \ll n$, $n \to \infty$ and $\delta \to \infty$. Importantly, note that the growth of δ need not be dependent on the growth of n. □

E. Proof of Theorem 4: Two Suspects

At last, in the case where $S = \{s_1, s_2\}$ contains only two suspect nodes, let d be the shortest path distance between s_1 and s_2 on the network G. We assume s_1 to be the rumor source s^* by symmetry, let $\mathcal{P} = \{v_0 = s_1, v_1, \cdots, v_d = s_2\}$ be the shortest path from s_1 to s_2, and define random variables Z_h be the numbers of nodes in subtrees $T_{v_h}^{s^*}$ $(1 \le h \le d)$. It is clear that $Z_h \ge Z_{h+1} + 1$ if $Z_h > 0$ for all $1 \le h \le d-1$.

In the following, we focus on the error detection probability, i.e., $\mathbf{P_e}(n) = 1 - \mathbf{P_c}(n)$. As a result, $Z_h > 0$ for all $1 \le h \le d$.

Since there are only two suspect nodes in S, $\mathbf{P_e}(n)$ can be written as

$$\begin{aligned}\mathbf{P_e}(n) &= \frac{1}{2}\cdot\sum_{R(s^*,G_n)=R(s_2,G_n)}\mathbf{P_G}\left[\bigcap_{h=1}^{d}(Z_h=z_h)\right]\\ &\quad+\sum_{R(s^*,G_n)<R(s_2,G_n)}\mathbf{P_G}\left[\bigcap_{h=1}^{d}(Z_h=z_h)\right],\end{aligned}\tag{53}$$

where $\mathbf{P_G}\left[\bigcap_{h=1}^{d}(Z_h=z_h)\right]$ is the probability of observing n infected nodes G_n with $Z_h = z_h$ nodes in subtree $T_{v_h}^{s^*}$ for all $1 \le h \le d$ on a regular tree with node degree δ; see its closed-form expression in (28).

As n is small, we can make use of numerical computation to calculate the error detection probability $\mathbf{P_e}(n)$ with (53). In the following, we will establish the argument of Theorem 4.

(1) Detection probability when $\delta = 2$

Proof of Theorem 4-i. By (4), we have $R(s^*, G_n) = \binom{n-1}{z_1}$ and $R(s_2, G_n) = \binom{n-1}{z_1-d}$. If $R(s^*, G_n) \le R(s_2, G_n)$, then $z_1 \ge (n+d+1)/2$. Notice that we only need to consider the distribution of Z_1. When $\delta = 2$ the distribution in (26) can be written as

$$\mathbf{P_G}\left[Z_1 = z_1\right] = \frac{(n-1)!}{z_1!(n-z_1-1)!}\frac{1}{2^{n-1}}.\tag{54}$$

By (53), the error detection probability is

$$\begin{aligned}\mathbf{P_e}(n) &= \frac{1}{2}\sum_{z_1=(n+d+1)/2}\mathbf{P_G}\left[Z_1=z_1\right]+\sum_{z_1>(n+d+1)/2}\mathbf{P_G}\left[Z_1=z_1\right],\\ &= \begin{cases}\dfrac{1}{2}-\dfrac{1}{2^n}\displaystyle\sum_{z_1=(n-d-1)/2}^{(n+d+1)/2}\binom{n-1}{z_1}, & (n-d)\text{ is odd};\\ \dfrac{1}{2}-\dfrac{1}{2^n}\displaystyle\sum_{z_1=(n-d)/2}^{(n+d-2)/2}\binom{n-1}{z_1}, & (n-d)\text{ is even}.\end{cases}\end{aligned}\tag{55}$$

Above, the detailed deduction is in Appendix-G. □

(2) Detection probability when $\delta = 3$

Proof of Theorem 4-ii. First of all, consider the case when $d = 1$. By (46), the error detection probability is

$$\begin{aligned}\mathbf{P_e}(n) &= 1-\mathbf{P_c}(n)\\ &= \frac{1}{4}-\frac{1}{4}\frac{1}{2\lfloor n/2\rfloor+1}.\end{aligned}\tag{56}$$

Now, consider the case when $d = 2$. The distribution in (28) can be written as

$$\mathbf{P_G}\left[\bigcap_{h=1}^{2}(Z_h=z_h)\right] = \frac{2(n-z_1)}{n(n+1)z_1}.\tag{57}$$

By (7), we have $R(s_2, G_n) = R(s^*, G_n)\frac{z_1}{n-z_1}\frac{z_2}{n-z_2}$. If $R(s^*, G_n) \le R(s_2, G_n)$, then $z_1 + z_2 \ge n$. By (53), the error detection probability as $n \to \infty$ is

$$\begin{aligned}\mathbf{P_e}(n) &= \frac{1}{2}\sum_{z_1+z_2=n}\mathbf{P_G}\left[\bigcap_{h=1}^{2}(Z_h=z_h)\right]\\ &\quad+\sum_{z_1+z_2>n}\mathbf{P_G}\left[\bigcap_{h=1}^{2}(Z_h=z_h)\right]\\ &\approx 0.114.\end{aligned}\tag{58}$$

Above, the detailed deduction is in Appendix-H. □

(3) Detection probability when $\delta > 3$

Proof of Theorem 4-iii. Consider the case when $d = 1$. By (49), the error detection probability is

$$\begin{aligned}\mathbf{P_e}(n) &= 1-\mathbf{P_c}(n)\\ &= 1-I_{1/2}\left(\frac{1}{\delta-2},\frac{\delta-1}{\delta-2}\right).\end{aligned}\tag{59}$$

as $n \to \infty$.

Furthermore, $I_{1/2}(0, 1) = 1$ (setting $\delta \to \infty$). Therefore,

$$\mathbf{P_e}(n) = 0,\tag{60}$$

for $n \to \infty$ and $\delta \to \infty$. □

(4) Increasing detection probability with distance

Proof of Theorem 4-iv. Equivalently, we prove that $\mathbf{P_e}(n)$ decreases with d. Without loss of generality we assume $Z_h > 0$ for all $1 \le h \le d$. The proof will be completed by showing that if $R(v_d, G_n) \ge R(s^*, G_n)$ then $R(v_{d-1}, G_n) > R(s^*, G_n)$ for all $d \ge 2$, which is to be verified by contradiction to the assumption $R(v_d, G_n) \ge R(s^*, G_n)$.

Suppose $R(v_{d-1}, G_n) \le R(s^*, G_n)$, then $Z_{d-1} \le n/2$. Otherwise, $Z_{d-1} > n/2$ and thus $Z_h > n/2$ for all $1 \le h \le d-1$; namely, $Z_h/(n-Z_h) > 1$ for all $1 \le h \le d-1$. Repeatedly using (7), we have

$$R(v_{d-1}, G_n) = R(s^*, G_n)\frac{Z_1}{n-Z_1}\frac{Z_2}{n-Z_2}\cdots\frac{Z_{d-1}}{n-Z_{d-1}},\tag{61}$$

which leads to the contradiction that $R(v_{d-1}, G_n) > R(s^*, G_n)$. As a result, we have $Z_{d-1} \le n/2$.

As $Z_{d-1} \le n/2$, then $Z_d < n/2$ and thus $Z_d/(n-Z_d) < 1$. As a result, we have

$$R(v_d, G_n) = R(v_{d-1}, G_n)\frac{Z_d}{n-Z_d} < R(v_{d-1}, G_n) \le R(s^*, G_n),\tag{62}$$

which is in contrary to the assumption $R(v_d, G_n) \ge R(s^*, G_n)$. □

IV. Experiments

In this section, we carry out simulation experiments under three scenarios to study our theoretical predictions. For verifying the asymptotic results, in each experiment run we let 1000 nodes be eventually infected by a rumor source node following the SI model, and use the estimator (3) to identify this source.

It is shown in Fig. 2 that the correct detection probability is increasing with the node degree δ, from virtually zero when $\delta = 2$ to 0.307 as δ exceeds 50. In Fig. 3, there are k connected suspect nodes, and the correct detection probability significantly exceeds $1/k$ when $\delta > 2$. Furthermore, reliable detection is achieved as δ grows large. In Fig. 4, there are two suspect nodes with a shortest path distance d, and the correct detection probability significantly exceeds $1/2$ when $\delta > 2$. Furthermore, the detection reliability increases as d increases, and reliable detection is also achieved whenever δ or d is sufficiently large.

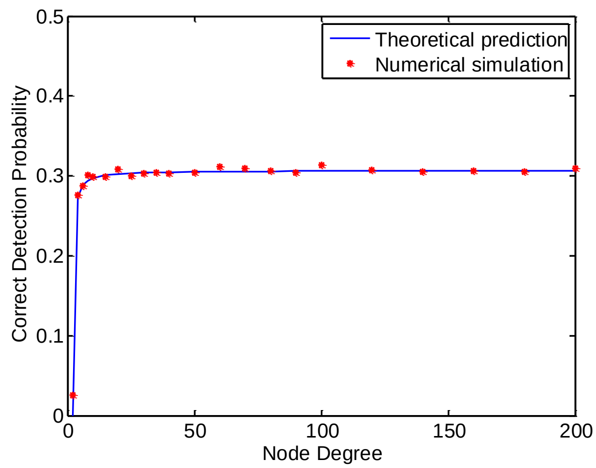

Fig. 6. Detection probability when $S = V$.

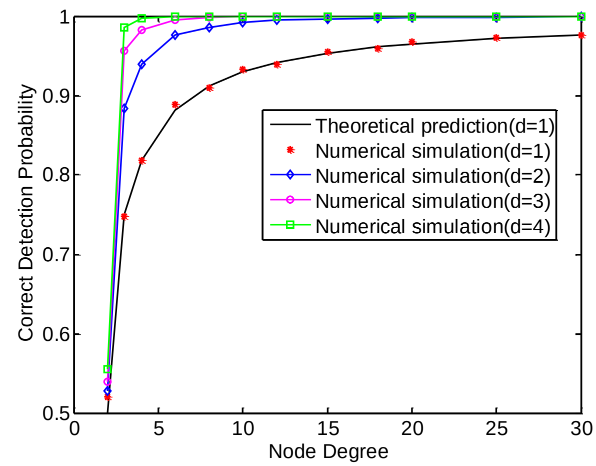

Fig. 8. Detection probability when S contains two suspect nodes.

V. Conlusions

In this paper, we have treated the problem of rooting out a single rumor source from a set of suspect nodes and constructed an MAP estimator as its solution. For regular tree-type networks, we have developed exact and asymptotic performance results of the source estimator under three scenarios. First, when all nodes are suspects, the correct detection probability is between 0.25 and 0.307 if the network is not linear ($\delta \neq 2$). Second, when the suspects form a small connected community, the correct detection probability significantly exceeds the *a priori* probability with $\delta \neq 2$, and reliable detection is achieved as the node degree becomes large. Third, with only two suspect nodes, the correct detection probability is at least 0.75 with $\delta \neq 2$, and it increases with the separation distance between the two suspects.

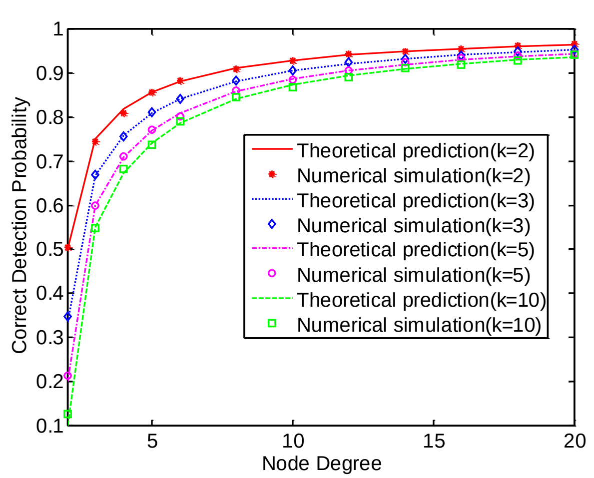

Fig. 7. Detection probability when S forms a connected subgraph of G.

Appendix

A. Proof of Lemma 5

Proof of Lemma 5. If $\max\{x_j, 1 \le j \le m\} < n/2$, then by Proposition 1 we know that s^* is the local rumor center w.r.t. the sub-neighborhood $N_l(s^*) = \{s_1^*, \dots, s_m^*\}$ of G_n. Again by Proposition 1, we have $R(u, G_n) < R(s^*, G_n)$ for all $u \in T_{s_j^*}^{s^*}$ and $1 \le j \le m$. Therefore, we can make sure to correctly identify s^* as the rumor source.

If $\max\{x_j, 1 \le j \le m\} = n/2$, then by Proposition 1 we know that s^* is the local rumor center w.r.t. the sub-neighborhood $N_l(s^*) = \{s_1^*, \dots, s_m^*\}$ of G_n. Again by Proposition 1, there is only a node $u \in N_l(s^*)$ such that $R(u, G_n) = R(s^*, G_n)$, and $R(v, G_n) < R(s^*, G_n)$ for all $v \in \{T_{s_j^*}^{s^*}, 1 \le j \le m\} \setminus \{u\}$. Therefore, the probability to correctly identify s^* as the rumor source is 1/2.

If $\max\{x_j, 1 \le j \le m\} > n/2$, then by Proposition 1 we know that s^* is not the local rumor center w.r.t. the sub-neighborhood $N_l(s^*) = \{s_1^*, \dots, s_m^*\}$ of G_n. Therefore, we can not identify s^* as the rumor source. □

B. Proof of Equation (32)

When n is even, we have

$$\begin{aligned}\mathbf{P_c}(n) &= p_{1/2} \cdot \sum_{\max\{x_1,x_2\}=n/2} \mathbf{P_G}\Big[\bigcap_{j=1}^{2}(X_j = x_j)\Big] \\ &\quad + p_1 \cdot \sum_{\max\{x_1,x_2\}<n/2} \mathbf{P_G}\Big[\bigcap_{j=1}^{2}(X_j = x_j)\Big] \\ &= \frac{1}{2} \cdot \mathbf{P_G}\Big[\frac{n-2}{2}, \frac{n}{2}\Big] + \frac{1}{2} \cdot \mathbf{P_G}\Big[\frac{n}{2}, \frac{n-2}{2}\Big],\end{aligned}$$

where we use the short form of $\mathbf{P_G}\Big[\bigcap_{j=1}^{2}(X_j = x_j)\Big]$, just as $\mathbf{P_G}\Big[x_j, 1 \le j \le 2\Big]$.

By (31), we have

$$\begin{aligned}\mathbf{P_c}(n) &= \frac{1}{2^n} \cdot \frac{(n-1)!}{((n-2)/2)!(n/2)!} + \frac{1}{2^n} \cdot \frac{(n-1)!}{(n/2)!((n-2)/2)!} \\ &= \frac{1}{2^{n-1}} \binom{n-1}{(n-2)/2}.\end{aligned}$$

When n is odd, we can similarly obtain

$$\mathbf{P_c}(n) = \frac{1}{2^{n-1}}\binom{n-1}{(n-1)/2}.$$

In summary, we have

$$\mathbf{P_c}(n) = \frac{1}{2^{n-1}}\binom{n-1}{\lfloor (n-1)/2 \rfloor}.$$

C. Proof of Equation (35)

When n is odd, we have

$$\begin{aligned}
\mathbf{P_c}(n) &= p_{1/2}\cdot \sum_{\max\{x_j,1\le j\le 3\}=n/2} \mathbf{P_G}\Big[\bigcap_{j=1}^{3}(X_j=x_j)\Big] \\
&\quad + p_1 \cdot \sum_{\max\{x_j,1\le j\le 3\}<n/2} \mathbf{P_G}\Big[\bigcap_{j=1}^{3}(X_j=x_j)\Big] \\
&= \mathbf{P_G}\Big[0,\frac{n-1}{2},\frac{n-1}{2}\Big] \\
&\quad + \mathbf{P_G}\Big[1,\frac{n-1}{2},\frac{n-3}{2}\Big] + \mathbf{P_G}\Big[1,\frac{n-3}{2},\frac{n-1}{2}\Big] \\
&\quad \vdots \\
&\quad + \mathbf{P_G}\Big[\frac{n-1}{2},\frac{n-1}{2},0\Big] + \cdots + \mathbf{P_G}\Big[\frac{n-1}{2},0,\frac{n-1}{2}\Big],
\end{aligned}$$

where we use the short form of $\mathbf{P_G}\Big[\bigcap_{j=1}^{3}(X_j = x_j)\Big]$, just as $\mathbf{P_G}\Big[x_j, 1\le j\le 3\Big]$.

By (34), we have

$$\begin{aligned}
\mathbf{P_c}(n) &= \frac{2}{n(n+1)}\cdot \sum_{x=0}^{x=(n-1)/2}(x+1) \\
&= \frac{1}{4}+\frac{3}{4n}.
\end{aligned}$$

When n is even, we can similarly obtain

$$\mathbf{P_c}(n) = \frac{1}{4}+\frac{3}{4}\frac{1}{n+1}.$$

In summary, we have

$$\mathbf{P_c}(n) = \frac{1}{4}+\frac{3}{4}\frac{1}{2\lfloor n/2\rfloor+1}.$$

D. Proof of Lemma 6

Proof of Lemma 6. Since the source s^* has m ($1\le m\le \delta$) neighbors in the suspect set S, by Proposition 1 and Lemma 5 we have

$$\begin{aligned}
\mathbf{P_c}(n|s^*) &\ge \mathbf{P_G}\Big[\bigcap_{i=1}^{m}E_i\Big] = 1-\mathbf{P_G}\Big[\bigcup_{i=1}^{m}E_i^c\Big] \\
&\overset{(a)}{\ge} 1-\sum_{i=1}^{m}\mathbf{P_G}\Big[E_i^c\Big] \overset{(b)}{=} 1-m\mathbf{P_G}\Big[E_1^c\Big].
\end{aligned}$$

Above, (a) is concluded by the union bound over events $E_1^c,\dots,E_m^c$, and (b) by symmetry.

Again by Proposition 1 and Lemma 5, we have

$$\begin{aligned}
\mathbf{P_c}(n|s^*) &\le \mathbf{P_G}\Big[\bigcap_{i=1}^{m}F_i\Big] = 1-\mathbf{P_G}\Big[\bigcup_{i=1}^{m}F_i^c\Big] \\
&\overset{(a)}{=} 1-\sum_{i=1}^{m}\mathbf{P_G}\Big[F_i^c\Big] \overset{(b)}{=} 1-m\mathbf{P_G}\Big[F_1^c\Big].
\end{aligned}$$

Above, (a) follows from the fact that events $F_1^c,\dots,F_m^c$ are disjoint since there is at most a subtree such that the number of nodes in it is more than $n/2$, and (b) from symmetry. □

E. Proof of Equation (42)

For $m=2$, by Appendix-B, we have

$$\mathbf{P_c}(n|s^*) = \frac{1}{2^{n-1}}\binom{n-1}{\lfloor (n-1)/2\rfloor}.$$

For $m=1$ and n is even, we have

$$\begin{aligned}
\mathbf{P_c}(n|s^*) &= p_{1/2}\cdot \sum_{x_1=n/2}\mathbf{P_G}\Big[\bigcap_{j=1}^{2}(X_j=x_j)\Big] \\
&\quad + p_1\cdot\sum_{x_1<n/2}\mathbf{P_G}\Big[\bigcap_{j=1}^{2}(X_j=x_j)\Big] \\
&= \frac{1}{2}\cdot\mathbf{P_G}\Big[\frac{n}{2},\frac{n-2}{2}\Big] \\
&\quad + \mathbf{P_G}\Big[\frac{n-2}{2},\frac{n}{2}\Big] + \cdots + \mathbf{P_G}\Big[0,n-1\Big],
\end{aligned}$$

where we use the short form of $\mathbf{P_G}\Big[\bigcap_{j=1}^{2}(X_j = x_j)\Big]$, just as $\mathbf{P_G}\Big[x_j, 1\le j\le 2\Big]$.

By (41), we have

$$\begin{aligned}
\mathbf{P_c}(n|s^*) &= \frac{1}{2^n}\cdot\binom{n-1}{n/2}+\frac{1}{2^{n-1}}\cdot\sum_{x=0}^{(n-2)/2}\binom{n-1}{x} \\
&= \frac{1}{2}+\frac{1}{2^n}\binom{n-1}{(n-2)/2}.
\end{aligned}$$

For $m=1$ and n is odd, we can similarly obtain

$$\mathbf{P_c}(n|s^*) = \frac{1}{2}+\frac{1}{2^n}\binom{n-1}{(n-1)/2}.$$

In summary, we have

$$\mathbf{P_c}(n|s^*) = \begin{cases}\frac{1}{2}+\frac{1}{2^n}\binom{n-1}{\lfloor (n-1)/2\rfloor}, m=1; \\ \frac{1}{2^{n-1}}\binom{n-1}{\lfloor (n-1)/2\rfloor}, m=2.\end{cases}$$

F. Proof of Equation (46)

Here, we only present the detailed proof of (46) when $m=2$. The case when $m=1$ can be deduced in a similar way, and the case when $m=3$ is the same as that in Appendix-C.

For $m=2$ and n is odd, we have

$$\begin{aligned}
\mathbf{P_c}(n|s^*) &= p_{1/2}\cdot\sum_{\max\{x_j,1\le j\le 2\}=n/2}\mathbf{P_G}\Big[\bigcap_{j=1}^{3}(X_j=x_j)\Big] \\
&\quad + p_1\cdot\sum_{\max\{x_j,1\le j\le 2\}<n/2}\mathbf{P_G}\Big[\bigcap_{j=1}^{3}(X_j=x_j)\Big] \\
&= \mathbf{P_G}\Big[0,\frac{n-1}{2},\frac{n-1}{2}\Big]+\cdots+\mathbf{P_G}\Big[0,0,n-1\Big] \\
&\quad + \mathbf{P_G}\Big[1,\frac{n-1}{2},\frac{n-3}{2}\Big]+\cdots+\mathbf{P_G}\Big[1,0,n-2\Big] \\
&\quad \vdots \\
&\quad + \mathbf{P_G}\Big[\frac{n-1}{2},\frac{n-1}{2},0\Big]+\cdots+\mathbf{P_G}\Big[\frac{n-1}{2},0,\frac{n-1}{2}\Big],
\end{aligned}$$

where we use the short form of $\mathbf{P_G}\Big[\bigcap_{j=1}^{3}(X_j = x_j)\Big]$, just as $\mathbf{P_G}\Big[x_j, 1 \le j \le 3\Big]$.

By (45), we have

$$\begin{aligned}\mathbf{P_c}(n|s^*) &= \frac{2}{n(n+1)} \cdot \sum_{x=0}^{x=(n-1)/2} \frac{n+1}{2} \\ &= \frac{1}{2} + \frac{1}{2n}.\end{aligned}$$

For $m = 2$ and n is even, we can similarly obtain

$$\mathbf{P_c}(n|s^*) = \frac{1}{2} + \frac{1}{2(n+1)}.$$

In summary, we have

$$\mathbf{P_c}(n|s^*) = \begin{cases} \frac{3}{4} + \frac{1}{4}\frac{1}{2\lfloor n/2 \rfloor + 1}, m = 1; \\ \frac{1}{2} + \frac{1}{2}\frac{1}{2\lfloor n/2 \rfloor + 1}, m = 2; \\ \frac{1}{4} + \frac{3}{4}\frac{1}{2\lfloor n/2 \rfloor + 1}, m = 3. \end{cases}$$

G. *Proof of Equation* (55)

Here, we only present the detailed proof of (55) when n is odd with even d. The other three cases when n is odd with odd d, n is even with even d, and n is even with odd d, can be deduced in a similar way.

Since $z_1 \ge (n+d+1)/2$ if and only if $R(s^*, G_n) \le R(s_2, G_n)$, thus by (54) we have

$$\begin{aligned}\mathbf{P_e}(n) &= \frac{1}{2}\sum_{z_1=(n+d+1)/2} \mathbf{P_G}\left[Z_1 = z_1\right] + \sum_{z_1>(n+d+1)/2} \mathbf{P_G}\left[Z_1 = z_1\right] \\ &= \frac{1}{2^n}\binom{n-1}{(n+d+1)/2} + \frac{1}{2^{n-1}}\sum_{z_1=(n+d+3)/2}^{n-1}\binom{n-1}{z_1} \\ &= \frac{1}{2} - \frac{1}{2^n}\sum_{z_1=(n-d-1)/2}^{(n+d+1)/2}\binom{n-1}{z_1}.\end{aligned}$$

H. *Proof of Equation* (58)

Here, we only present the detailed proof of (58) when n is even. The other case when n is odd can be deduced in a similar way.

Since $z_1 + z_2 \ge n$ if and only if $R(s^*, G_n) \le R(s_2, G_n)$. Besides, $z_1 \ge z_2 + 1$, thus $z_1 \ge (n+1)/2$. Therefore, by (57) we have

$$\begin{aligned}\mathbf{P_e}(n) &= \frac{1}{2}\sum_{z_1+z_2=n} \mathbf{P_G}\left[\bigcap_{h=1}^{2}(Z_h = z_h)\right] \\ &+ \sum_{z_1+z_2>n} \mathbf{P_G}\left[\bigcap_{h=1}^{2}(Z_h = z_h)\right] \\ &= \frac{1}{2}\mathbf{P_G}\Big[\frac{n}{2}+1, \frac{n}{2}-1\Big] + \cdots + \frac{1}{2}\mathbf{P_G}\Big[n-1, 1\Big] \\ &+ \mathbf{P_G}\Big[\frac{n}{2}+1, \frac{n}{2}\Big] \\ &+ \mathbf{P_G}\Big[\frac{n}{2}+2, \frac{n}{2}-1\Big] + \cdots + \mathbf{P_G}\Big[\frac{n}{2}+2, \frac{n}{2}+1\Big] \\ &\vdots \\ &+ \mathbf{P_G}\Big[n-1, 2\Big] + \cdots + \mathbf{P_G}\Big[n-1, n-2\Big] \\ &= \frac{2}{n(n+1)}\frac{n/2-1}{n/2+1}\left(\frac{1}{2}+1\right) + \frac{2}{n(n+1)}\frac{n/2-2}{n/2+2}\left(\frac{1}{2}+3\right) \\ &\quad + \cdots + \frac{2}{n(n+1)}\frac{n/2-(n-2)/2}{n/2+(n-2)/2}\left(\frac{1}{2}+n-3\right),\end{aligned}$$

where we use the short form of $\mathbf{P_G}\Big[\bigcap_{h=1}^{3}(Z_h = z_h)\Big]$, just as $\mathbf{P_G}\Big[z_h, 1 \le h \le 3\Big]$.

As $n \to \infty$, we have

$$\lim_{n\to\infty} \mathbf{P_e}(n) \approx 0.114.$$

Acknowledgment

The work of W. Dong and W. Zhang has been supported by National Basic Research Program of China (973 Program) through grant 2012CB316004, and by the 100 Talents Program of Chinese Academy of Sciences. The work of C. W. Tan has been supported by the Research Grants Council of Hong Kong under Project No. RGC CityU 125212.

References

[1] D. Shah and T. Zaman, "Rumors in a network: who's the culprit?" *IEEE Trans. Inform. Theory*, vol. 57, no. 8, pp. 5163–5181, Aug. 2011.

[2] ——, "Rumor centrality: a universal source detector," in *SIGMETRICS'12: Proc. of 12th ACM SIGMETRICS/PERFORMANCE joint int. conf. on measurement and modeling of computer systems*, 2012, pp. 199–210.

[3] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications*. Cambridge Univ. Press, 1994.

[4] D. Kempe, J. Kleinberg, and E. Tardos, "Maximizing the spread of influence through a social network," in *SIGKDD'03: Proc. of 9th ACM SIGKDD int. conf. on knowledge discovery and data mining*, 2003, pp. 137–146.

[5] N. T. J. Bailey, *The Mathematical Theory of Infectious Diseases and its Applications*, 2nd ed. Griffin, 1975.

[6] C. Moore and M. E. J. Newman, "Epidemics and percolation in small-world networks," *Phys. Rev. E*, vol. 61, no. 5, pp. 5678–5682, May 2000.

[7] R. Pastor-Satorras and A. Vespignani, "Epidemic spreading in scale-free networks," *Phys. Rev. Lett.*, vol. 86, no. 14, pp. 3200–3203, Apr. 2001.

[8] M. E. J. Newman, "Spread of epidemic disease on networks," *Phys. Rev. E*, vol. 66, no. 1, p. 016128, Jul. 2002.

[9] P. G. Lind, L. R. da Silva, J. José S. Andrade, and H. J. Herrmann, "Spreading gossip in social networks," *Phys. Rev. E*, vol. 76, no. 3, p. 036117, Sep. 2007.

[10] A. Ganesh, L. Massoulie, and D. Towsley, "The effect of network topology on the spread of epidemics," in *INFOCOM'05: Proc. of 24th annu. joint conf. of the IEEE computer and communications Societies*, vol. 2, 2005, pp. 1455–1466.

[11] G. Streftaris and G. J. Gibson, "Statistical inference for stochastic epidemic models," in *Proc. of 17th Int. Workshop on Stat. Model.*, 2002, pp. 609–616.

[12] N. Demiris and P. D. O'Neill, "Bayesian inference for stochastic multitype epidemics in structured populations via random graphs," *J. of the Royal Stat. Soc. (B)*, vol. 67, no. 5, pp. 731–745, 2005.

[13] H. Okamura, K. Tateishi, and T. Dohi, "Statistical inference of computer virus propagation using non-homogeneous poisson processes," in *ISSRE'07: Proc. of 18th IEEE Int. Symp. on Software Reliability*, vol. 5, Nov. 2007, pp. 149–158.

[14] M. Gomez-Rodriguez, J. Leskovec, and A. Krause, "Inferring networks of diffusion and influence," in *SIGKDD'10: Proc. of 16th ACM SIGKDD int. conf. on knowledge discovery and data mining*, 2010, pp. 1019–1028.

[15] W. Chen, Y. Wang, and S. Yang, "Efficient influence maximization in social networks," in *SIGKDD'09: Proc. of 15th ACM SIGKDD int. conf. on knowledge discovery and data mining*, 2009, pp. 199–208.

[16] W. Dong, W. Zhang, and G. Wei, "Extracting influential information sources for gossiping," in *Allerton'12: Proc. of 50th annu. Allerton conf. on communication, control and computing*, 2012.

[17] D. Shah and T. Zaman, "Finding rumor sources on random graphs," *arXiv-v1110.6230*, 2011.

[18] W. Luo, W. P. Tay, and M. Leng, "Identifying infection sources and regions in large networks," *arXiv-v1204.0354*, 2012.

[19] K. Zhu and L. Ying, "Information source detection in the SIR model: a sample path based approach," *arXiv-v1206.5421*, 2012.

[20] P. C. Pinto, P. Thiran, and M. Vetterli, "Locating the source of diffusion in large-scale networks," *Phys. Rev. Lett.*, vol. 109, no. 6, p. 068702, Aug. 2012.

[21] N. L. Johnson and S. Kotz, *Urn Models and Their Application: An Approach to Modern Discrete Probability Theory*. John Wiley & Sons, 1977.

[22] G. Brightwell and P. Winkler, "Counting linear extensions is #P-complete," in *STOC'91: Proc. of 23rd annu. ACM symp. on theory of computing*, 1991, pp. 175–181.